Predicting the neutral hydrogen content of galaxies from optical data using machine learning

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **479,** 4509–4525 (2018)
Advance Access publication 2018 July 5

doi:10.1093/mnras/sty1777

Predicting the neutral hydrogen content of galaxies from optical data using machine learning

Mika Rafieferantsoa,[1,2,3]★ Sambatra Andrianomena[1,4]★ and Romeel Davé[1,3,5]

[1]*University of the Western Cape, Bellville, Cape Town 7535, South Africa*
[2]*Max-Planck-Institüt für Astrophysik, D-85748 Garching, Germany*
[3]*South African Astronomical Observatory, Observatory, Cape Town 7925, South Africa*
[4]*SKA South Africa, 3rd Floor, The Park, Park Road, Pinelands, 7405, South Africa*
[5]*Institute for Astronomy, Royal Observatory, Edinburgh EH9 3HJ, UK*

Accepted 2018 July 1. Received 2018 May 28; in original form 2018 March 22

ABSTRACT

We develop a machine learning-based framework to predict the H I content of galaxies from optical photometry and environmental parameters. We train the algorithm on $z = 0$–2 outputs from the MUFASA cosmological hydrodynamic simulation, which includes star formation, feedback, and a heuristic model to quench massive galaxies that yields a reasonable match to a range of survey data including H I. We employ a variety of machine learning methods (regressors), and quantify their performance using the slope of the predicted versus true relation, its root mean square error (RMSE), and Pearson correlation coefficient (**r**). Training on only Sloan Digital Sky Survey photometry, all regressors give $\mathbf{r} > 0.8$ and RMSE ~ 0.3 at $z = 0$, led by random forests with $\mathbf{r} = 0.91$, and a deep neural network (DNN) with comparable accuracy ($\mathbf{r} = 0.9$). Adding near-IR photometry improves all regressors. All regressors perform worse with redshift, particularly at $z \gtrsim 1$. Slope values are generally sub-linear, so that we overpredict H I in H I-poor galaxies and underpredict H I rich, because the regressors do not fully capture the scatter in the data. We test our framework on REsolved Spectroscopy Of a Local VolumE (RESOLVE) and Arecibo Legacy Fast ALFA (ALFALFA) survey data. Training on a subset of the observations, we find that our machine learning method can reasonably predict H I richnesses in the remaining data (RMSE ~ 0.28 for RESOLVE and $\sim$0.25 for ALFALFA). Training on mock data from MUFASA to predict observed data is worse (RMSE ~ 0.45 for RESOLVE and 0.31 for ALFALFA), with DNN well outperforming other regressors. Our method will be useful for making galaxy-by-galaxy survey predictions and incompleteness corrections for upcoming H I 21 cm surveys on Square Kilometre Array precursors such as MeerKAT, over regions where photometry is already available.

Key words: methods: numerical – galaxies: evolution – galaxies: statistics.

1 INTRODUCTION

One of the most important science goals of the Square Kilometre Array (SKA) project is to provide us more insights into the growth and fuelling of galaxies. A particular focus is on the evolution of their atomic neutral hydrogen, or H I content, which constitutes a major part of the gas content of galaxies, as traced by 21 cm radio emission. H I gas represents the dense gas reservoir that will eventually form stars after passing through a molecular phase, and is thus a key and so far underexplored aspect of the baryon cycle governing galaxy evolution (Somerville & Davé 2015). Hence, upcoming surveys with SKA precursors MeerKAT and ASKAP aim to expand the depth and area of 21 cm surveys out to $z \sim 1$, with the SKA potentially reaching even higher redshifts.

Much work has been done on studying the H I content of galaxies in the nearby Universe. The Arecibo Legacy Fast ALFA (ALFALFA; Giovanelli et al. 2005) blindly observed about 7000 deg^2 of the Arecibo sky and was complete in 2012. It has enabled a precise study of the distribution of galaxies in the local Universe based on their HI mass. For instance, Jones et al. (2016) studied the environmental effects on the H I content of galaxies using the ALFALFA survey $\alpha.70$ (70 per cent of the final data). They found a shift of the Schechter function knee towards higher value in higher density environments. Due to ALFALFA's high positional accuracy of <20 arcsec, they could explore the optical counterparts and extend the understanding of the stellar mass growth based on

★ E-mail: rafieferantsoamika@gmail.com (MR); andrianomena@gmail.com (SA)

H I content. The GALEX Arecibo SDSS Survey (GASS; Catinella et al. 2010) used a complementary approach by selecting ∼800 $L*$ galaxies from the Sloan Digital Sky Survey (SDSS; York et al. 2000) and observed their H I-line spectra until detection. Catinella et al. (2010) found that the *detected* (60 per cent of the 20 per cent observed) H I richness ($M_{\rm H\,I}/M_*$) does not go below 40 per cent even for the highest stellar masses explored (∼10^{11} $M_\odot$). Using the full GASS data set, Catinella et al. (2013) found an environment dependence of the gas fraction, such that galaxies in higher host halo masses have lower H I than those in less dense environments, confirming the idea that galaxy gas content and environment are tightly connected. The REsolved Spectroscopy Of a Local VolumE (RESOLVE; Kannappan et al. 2011) survey adopted yet another approach by observing ∼1500 galaxies with ranges of stellar and gas masses within a volume-limited 53 000 Mpc^3 in the nearby Universe. Stark et al. (2016) used the RESOLVE data, targeting an area within the SDSS redshift survey, and found that decreasing H I richness in galaxies is related to increasing host halo mass for a given stellar content. These data set the stage for explorations to lower masses and higher redshifts to be achieved with next-generation surveys.

Theoretical studies on the evolution of H I content of galaxies have also been expanding. Cunnama et al. (2014) predicted from the Galaxies-Intergalactic Medium Interaction Calculation suites of hydrodynamical simulations (Crain et al. 2009), a tight dependence of galaxies' H I column density and environment: Galaxies in groups possess extended H I radial profiles compared to field galaxies. The extended radial profiles originate from the ram pressure redistribution which they found to dominate over the gravitational restoring forces. Although their findings are physically grounded, disentangling such processes remain a challenge for observers. Related results were found using a different galaxy formation model from Davé et al. (2013), where Rafieferantsoa et al. (2015) found a faster depletion of H I content once galaxies fall in a more massive haloes. The specific star formation rate of those galaxies also decreases but at rate less than that of the H I, indicating gas stripping from the outskirt of the galaxies inwards. Quilis, Planelles & Ricciardelli (2017) studied the effects of ram pressure stripping. They used a cosmological simulated box to select a sample of galaxies residing in clusters to do their analysis. They found that galaxies below 10^{10} $M_\odot$ in stellar mass are often located at the outskirts of the clusters and have high eccentricity. Their interactions with the environment are more violent resulting in faster change of the gas contents and morphologies of the galaxies. More massive galaxies are situated closer to the cluster centres, and the gas removal is less dominant. The major change in those galaxies is caused by inflowing gas from the intercluster medium. Using the MUFASA data (Davé, Thompson & Hopkins 2016), Rafieferantsoa & Davé (2018) found a weak but extended galactic conformity in H I richness for galaxy members of low-mass haloes. Bigger host-halo galaxies tend to have stronger but less extended conformity. These studies demonstrate that the H I content of galaxies is impacted by their environment, but the exact nature of that dependence is not entirely clear.

Hence, observational surveys suggest that understanding the baryon cycle requires precise measurements of the H I content of the galaxies, which at times might be affected by observational artefacts. Theoretical works, on the other hand, predict physical results that are currently difficult to observe, which argues for larger and deeper H I surveys to improve our current understanding of the evolution of gas content and hence galaxy growth overall.

Although considerable efforts have gone into studying the gas phase properties of galaxies with the help of the currently available H I data, e.g. ALFALFA and RESOLVE, the understanding of H I evolution still lags behind the understanding of their stellar components. The main reason is that photometric data can be directly related to the stellar population of galaxies, and such optical and near-infrared data are currently technologically able to reach deeper levels than radio data. For the promise of multiwavelengths surveys to be fully realized into the radio regime, it is important to be able to relate gas and stellar properties accurately. However, this is not straightforward. There have been attempts that have been proposed to estimate gas-phase properties of galaxies from their stellar masses obtained from spectral energy distribution (SED) fitting to photometrical properties. For instance, Kannappan (2004) found a correlation between $u-K$ colours and H I richness that they dubbed *photometric gas fractions*. The correlation was shown to be valid for galaxies with atomic gas fraction ranging from 1 per cent to 10× the stellar masses. Zhang et al. (2009) developed a similar method by using the i-band surface brightness and the $g-r$ colour to estimate the H I richness of the galaxies. They found a tighter scatter compared to previous estimations. The H I scaling relations found by Zhang et al. (2009) were further improved upon by Wang et al. (2013) by introducing a form of correction to account for the fact that H I-rich galaxies have more active star formation on the outer discs (bluer) (see Wang et al. 2011). Still with the standard approach by first establishing correlation between the gas fraction and other galaxy properties, Catinella et al. (2010) prescribed another relation $\log_{10}(M_{\rm H\,I}/M_*) = -0.332\log_{10}(\mu_*) - 0.240(NUV-r) + 2.856$ that was also tested by Wang et al. (2015) with their samples to estimate the gas fraction as a function of stellar mass surface density (μ_*) and observed $NUV-r$ colour. Teimoorinia, Ellison & Patton (2017), in a work most similar to ours, used machine learning (ML) that was trained on ALFALFA data to predict the H I content of half a million SDSS galaxies. In addition to direct photometric data, they considered 14 other derived galaxy properties as input parameters, and attained their best performance of only ∼0.2 dex off the observed quantities. From these studies, it is clear that developing ways to connect optical/NIR information with H I is an important task, which affords many applications such as to estimate the H I content of certain galaxies based solely on their available photometry information, to enable larger statistics, and to assess incompleteness in surveys.

In this work, we explore a more general approach compared to previous studies by investigating the feasibility of predicting the H I richness of galaxies from directly observed optical properties of galaxies, particularly the photometric magnitudes and environmental quantities, using ML. The main idea is that ML can synthesize all the photometric data in order to optimally predict H I, rather than trying to isolate particular combinations that work best. The advantage of using ML techniques is mainly the capability of the model to learn peculiar aspects human might have overlooked, with the downside that such a method does not provide a direct physical interpretation of the result. The choice of only using directly observed quantities avoids introducing systematic uncertainties arising from the estimation of derived quantities such as stellar mass and star formation rate. For this paper, we focus on galaxies having at least some H I content; future work will explore identifying that galaxies are gas free.

Another key difference with respect to previous works is that we test the efficacy of using cosmological hydrodynamic simulations as a training set. The advantage of this is that, if our model is sufficiently good representative of the real Universe, then it can be used to explore regimes where HI data do not currently exist, such as those at higher redshifts. This provides a more physically motivated

prediction of H I content than using locally calibrated relations. Furthermore, by using simulated galaxies to train and calibrate the method, connections can be made between the obtained correlations and the underlying physics, at least within the context of the given model.

Our best ML algorithms, random forest (RF) and deep learning, are able to predict the H I richness of simulated galaxies to within <0.3 dex from their real values using only the photometric properties of the simulated galaxies. Testing this on RESOLVE and ALFALFA survey data, the prediction of the observed data from simulation-trained models yield less precise results. Generally, RF is our optimal ML algorithm, but the neural network's performance becomes better when observational data are used.

Our method has numerous applications. Current data as well as future surveys will benefit from this method by providing ways to more accurately correct observations for incompleteness and confusion. For instance, the upcoming Looking At the Distant Universe with the MeerKAT Array (LADUMA; Holwerda, Blyth & Baker 2012) survey aims to directly detect and use different techniques to stack multiple objects to be able to measure H I fluxes out to $z \sim 1$ for the first time, to enable a deeper understanding of the fuelling processes of galaxies and study the cosmic evolution of their H I content. But at higher redshift, confusion can become dominant especially when sources are located in groups. Meanwhile, ASKAP H I All-Sky Survey (WALLABY) that will cover two third of the sky will probe H I gas of 6×10^5 galaxies up to $z = 0.26$; DINGO, up to $z = 0.43$, will probe about 10^5 galaxies within 4×10^7 Mpc^3 cosmological volume (Duffy et al. 2012). These H I surveys will provide a wealth of information on galaxy evolution, but it is important to be able to accurately measure and understand the observations, which is where our method can provide insights.

Section 2 briefly reviews the MUFASA simulation used for this work. The approach we use in this study is detailed in Section 3, and we present the techniques utilized in order to achieve our goal in Section 4. Section 5 presents our findings, and Section 6 shows a preliminary application of our method. We expand on the limitations of our method in Section 7 and finally conclude in Section 8.

2 SIMULATIONS

2.1 Galaxy formation models: MUFASA

For our training set we make use of the outputs of the MUFASA simulation model, which is fully described in Davé et al. (2016). We only present the key prescriptions in the model that are particularly relevant for this work.

MUFASA is implemented in the GIZMO cosmological hydrodynamics code, including a tree-particle-mesh gravity code based on GADGET (Springel 2005), topped with a meshless finite mass hydrodynamic algorithm (Hopkins 2015). The model uses radiative cooling and heating implemented with the GRACKLE 2.1 library.[1] Star formation follows a Schmidt (1959) law, based on a subgrid prescription that determines the molecular gas content of each gas particles (Krumholz & Gnedin 2011), and occurs only in gas elements above a hydrogen number density threshold of $n_{\rm H} > 0.13\,{\rm cm}^{-3}$.

MUFASA uses a kinetic gas outflow prescription to model star-formation driven winds, following scalings from the Feedback in Realistic Environments (Muratov et al. 2015) zoom simulations. MUFASA also contains a heuristic prescription for star formation quenching, whereby it heats the gas volume elements within a host halo that are above a halo mass threshold of $M_{\rm halo} > (1 + 0.48z) \times 10^{12}\,{\rm M}_\odot$ (Gabor & Davé 2015; Mitra, Davé & Finlator 2015). This model is intended to mimic radio mode feedback from active galactic nuclei (Croton et al. 2006) in massive haloes.

2.2 Galaxy sample

The galaxy sample used for our analysis is obtained by simulating a cube of $50\,h^{-1}$ Mpc on a side with 512^3 dark matter particles and 512^3 gas volume elements. The initial conditions are generated at redshift $z = 249$ using MUSIC (Hahn & Abel 2011) with Planck Collaboration XIII (2016) concordant cosmological parameters, namely $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.048$, $H_0 = 68\,{\rm km\,s^{-1}\,Mpc^{-1}}$, $\sigma_8 = 0.82$, and $n_s = 0.97$.

MUFASA evolves these initial conditions to $z = 0$, outputting 135 snapshots. For each snapshot, we identify galaxies, with SKID[2] (Kereš et al. 2005) as gravitationally bound collections of stars and star-forming gas. In our analysis, we will only use $z \leq 2$ sample, which, in total, is made of 50 snapshots. Each snapshot contain typically around 8000 resolved galaxies (>64 star particle masses or $M_* > 1.16 \times 10^9\,{\rm M}_\odot$).

2.3 Galaxy properties

Our simulated galaxy properties are calculated with a modified version of CAESAR,[3] which is an add-on package for the YT simulation analysis suite. The stellar mass of a galaxy, or M_*, is the total mass of the stellar particles within it. The atomic neutral hydrogen content, $M_{\rm H\,I}$, of the galaxy is the summation of all H I from the gas particles. For each gas volume element, we account for the self-shielding from the metagalactic UV background radiation, by using a fitting formula for the effective optically thin photoionization rate as a function of density (Rahmati et al. 2013). The galaxy peculiar velocity $v_{\rm gal}$ is the one-dimensional (1D) mass-weighted average of all the particle velocities contained in it, along each of the (x, y, z) directions. We use the projected nearest neighbour density Σ_3 to quantify the galaxy environment such that

$$\Sigma_3 = \frac{3}{\pi R_3^2}, \tag{1}$$

where R_3 is the distance of the galaxy to its third closest neighbour, projected along the line of sight.

The magnitudes of the galaxies are obtained using the LOSER[4] (see Davé, Rafieferantsoa & Thompson 2017b, for a fuller description) package (not CAESAR) but still using the groups identified by SKID. We first use the Flexible Stellar Population Synthesis (FSPS; Conroy & Gunn 2010) library to derive the stellar spectra of each star particle based on its age and metallicity, summing to obtain the stellar spectrum for that galaxy. Every stellar spectrum is attenuated by the line-of-sight dust extinction obtained by scaling the metal column density along the given line of sight; this results in each of six lines of sight ($\pm x$, $\pm y$, $\pm z$) having different extinction and thus different spectra. We obtain all magnitudes by applying the appropriate filters. We computed (u,g,r,i,z) SDSS magnitudes, (U,V) Johnson magnitudes, *NUV* GALEX magnitude, and the (J, H, K_s) 2MASS magnitudes.

[1] https://grackle.readthedocs.io/en/grackle-2.1/genindex.html

[2] http://www-hpcc.astro.washington.edu/tools/skid.html

[3] https://bitbucket.org/laskalam/caesar

[4] Line Of Sight Extinction by Ray-tracing (https://bitbucket.org/romeeld/closer).

3 ML SET-UP

The goal is to predict the H I richness ($M_{\rm H\,I}/M_*$) from other properties of a given galaxy. We use the supervised learning paradigm that consists of training the algorithm to estimate the desired label when fed with a corresponding input. Through a learning process, the best model parameters that minimize a defined cost function, which we choose to be the mean squared errors (MSEs), are solved. Sets of training data sets drawn from our simulated sample are used to train our learners to predict the target ($M_{\rm H\,I}/M_*$) from the features $\{u, g, r, i, z, U, V, J, H, K_s, \Sigma_3, v_{\rm gal}\}$ of our galaxies.

It is noted that $v_{\rm gal}$ indicates line-of-sight velocity, and our models will predict the H I richness ($M_{\rm H\,I}/M_*$) of the galaxies rather than their $M_{\rm H\,I}$ due to the less constrained correlation between the latter and the galaxy stellar masses. In addition, we take the logarithmic values of the target due to its large dynamic range that can cause the learning process to fail. First of its series, this work focuses only on the prediction of the H I richness of H I rich galaxies and to do so, we only select galaxies with $M_{\rm H\,I}/M_* > 10^{-2}$, which decreases the size of our sample. To counteract, we increase our data by calculating the galaxy properties along all the six projections axis of the simulated cubical box, resulting in 6 × more data for our analysis.

We assume we have all photometric magnitudes for all available bands, covering a wide range of spectrum including SDSS magnitudes, Johnson magnitudes and 2MASS magnitudes, which we can compute from our simulated galaxies. Although this scenario is ideal for our analysis, it is not so realistic. We can expect observed galaxies to only have $\{u,g,r,i,z\}$ magnitudes at best. To this regard, we examine different possibilities in our analysis. All the set-ups considered in this work are listed in Table 1, where `color indices` denotes all possible pairwise combination (e.g. $g - r$) of all the magnitudes in the surveys considered in one set-up.

We train our model in two different ways. First is the '*f*-training', which considers all the galaxies from all the $z \leq 2$ outputs (with *f* leading the set-up names, see the first column of Table 1). Second is the '*z*-training', in which we build a regressor at each redshift bin (with *z* leading the set-up names). In both approaches, we randomly choose 75 per cent of the data as the training set and 25 per cent as testing set. We do the training 10 times with 10 different random batches to get the uncertainty of our results.[5]

To this end, we make use of six different ML techniques that we describe in the following.

4 ML ALGORITHMS

We use `TensorFlow` to build the DNN model and `scikit-learn` (Pedregosa et al. 2011) package for the remaining methods.

4.1 Linear regression

Linear regression (LR) model [along with *k*-nearest neighbour (*k*-NN), see Section 4.3] is the simplest amongst those we use in this work. Its simplicity, hence its great speed during training, provides quick insights into the relationship between the features ($\mathbf{x}$) and the corresponding target (y). The latter is defined as a linear combination of all the features, $y = \mathbf{w} \cdot \mathbf{x}$, and the idea consists of finding the weights $\mathbf{w}$ that minimize the MSE

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}\cdot\mathbf{x}_n - y_n)^2. \tag{2}$$

Here, the bias is absorbed into the weights $\mathbf{w}$.

4.2 Ensemble learning methods: RF and Gradient Boosting (GRAD)

To understand both RF and GRAD algorithms, one needs to first look at their base estimators, the Decision trees (Hastie, Tibshirani & Friedman 2009), which will be clarified below.

In a simple 1D problem, we assume a data set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ of length N ($(x, y) \in \mathbb{R} \times \mathbb{R}$). The first step of the algorithm is to split the training set at a split point s that minimizes the cost function

$$J = \min_{c_1}\left\{\sum_{x_i\in R_1(s)}(y_i - c_1)^2\right\} + \min_{c_2}\left\{\sum_{x_i\in R_2(s)}(y_i - c_2)^2\right\}, \tag{3}$$

where $R_1 = \{x_i | x_i \leq s\}$ and $R_2 = \{x_i | x_i > s\}$ are the two regions (also called nodes) resulting from the split. The values c_1 and c_2 that minimize each term in equation (3) are simply the averages of the labels y_i in R_1 and R_2, respectively, i.e.

$$c_1 = \frac{1}{m_1}\sum_{x_i\in R_1(s)} y_i,$$
$$c_2 = \frac{1}{m_2}\sum_{x_i\in R_2(s)} y_i, \tag{4}$$

where m_1 and m_2 are the number of inputs x_i found in R_1 and R_2, respectively. To grow the tree, each resulting node from the root is further split recursively (known as greedy algorithm) until a fixed maximum depth (or size) of the tree is reached. The nodes at the bottom of the tree are called the leaf nodes. To predict a new label $y_{\rm new}$ from a new input $x_{\rm new}$, one simply walks through the tree from the root to reach a leaf node that then estimates $y_{\rm new}$ by averaging the corresponding labels y_i of the inputs x_i within it according to[6]

$$\hat{y}_{\rm new} = \frac{1}{m}\sum_{x_i\in\mathcal{L}} y_i, \tag{5}$$

where $\mathcal{L}$ indicates the leaf node and m the number of points x_i within it. Decision trees are prone to overfitting but there exist various techniques of regularization.

RF (Breiman 2001), known to be a powerful ML algorithm, is composed of a given number[7] of decision trees (base estimators) that are individually trained with a random subset of the data set. To do a prediction, RF simply averages the predictions of its decision trees.

Another well-known ensemble learning model that we use is gradient boosting (Friedman 2000). Its base learner is also a decision tree but instead of simply aggregating the predictions of its regressors like in the case of RF, the training is carried out in a sequence. Except for the first regressor, which is trained with the data set, each next regressor in the sequence[8] fits the residual errors of its predecessor and so on. The resulting estimator is then of the following

[5] At each iteration, the data set is randomly shuffled and new batches of training and test sets are generated.

[6] Similar to equation (4).

[7] Which is amongst the hyper-parameters of the model.

[8] This is set by the number of the base estimators.

Table 1. List of all the set-ups that are considered in the analysis. For easy reference, each set-up has been given a name.

Name	Surveys	Features	Target	Description
fSMg	SDSS	$u, g, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Redshift information not required
fSClr	SDSS	`color indices`, $v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Redshift information not required
fSCmb	SDSS	`color indices`, $u, g, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Redshift information not required
fAMg	SDSS+Johnson+2MASS	$H, J, Ks, U, V, u, g, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Redshift information not required
fAClr	SDSS+Johnson+2MASS	`color indices`, $v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Redshift information not required
zSMg	SDSS	$u, g, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Prediction at a given redshift bin
zSClr	SDSS	`color indices`, $v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Prediction at a given redshift bin
zSCmb	SDSS	`color indices`, $u, g, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Prediction at a given redshift bin
zAMg	SDSS+Johnson+2MASS	$H, J, Ks, U, V, u, r, r, i, z, v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Prediction at a given redshift bin
zAClr	SDSS+Johnson+2MASS	`color indices`, $v_{\rm gal}, \Sigma_3$	$\log(M_{\rm H\,I}/M_*)$	Prediction at a given redshift bin

form:

$$\mathcal{E}(x) = \mathcal{E}_1(x) + \sum_{i=2}^{N} \gamma_i e_i(\epsilon_i), \tag{6}$$

where $\mathcal{E}_1(x)$ is the first estimator, ϵ_i the residual errors from the $i - 1^{\rm th}$ learner used as inputs of the ith learner to fit a predictor e_i and γ_i is a coupling parameter that is optimized such that the error from the combined system at each iteration ($i.e.$ $\mathcal{E}_{i+1}(x) = \mathcal{E}_1(x) + \sum_{k=2}^{i} \gamma_k e_k(\epsilon_k)$) is minimized. N is the number of base regressors (equal to the number of iteration) that form the ensemble.

4.3 *k*-Nearest neighbour

k-NN (Altman 1992) is a flexible non-parametric regression algorithm. Considering a set of instances $\mathbf{x}_n$ (in general, $\mathbf{x}_n \in \mathbb{R}^d$, but for the sake of simplicity, we let $\mathbf{x}_n \in \mathbb{R}$) with their corresponding label y_n ($y_n \in \mathbb{R}$), to predict a new label $y_{\rm new}$ given a new instance $\mathbf{x}_{\rm new}$, the estimate of $y_{\rm new}$ is simply the weighted average of targets of the k-closest neighbours of $\mathbf{x}_{\rm new}$. The principle is generalized for d-dimensions in feature space.

4.4 Support vector machine

Given a set of training data consisting of examples $\mathbf{x}_n$ ($\mathbf{x}_n \in \mathbb{R}^d$) and their labels y_n ($y_n \in \mathbb{R}$), the method aims at finding a linear function of the form $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. This can be seen as a convex optimization that seeks to

(i) minimize $\frac{1}{2}\mathbf{w}^T\mathbf{w}$,
subject to the constraint $|y_n - (\mathbf{w} \cdot \mathbf{x}_n + b)| \leq \epsilon$,

where ϵ denotes the residuals between estimates and the desired outputs. To deal with otherwise intractable optimization problem, Vapnik (1995) introduced some slack variables ξ_n^-, ξ_n^+ such that it now aims at

(i) minimizing $\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\xi_n^- + \xi_n^+)$
subject to

$$\text{the constraints } \begin{cases} y_n - (\mathbf{w} \cdot \mathbf{x}_n + b) \leq \epsilon + \xi_n^-, \\ \mathbf{w} \cdot \mathbf{x}_n + b - y_n \leq \epsilon + \xi_n^+, \\ \xi_n^-, \xi_n^+ \geq 0, \end{cases} \tag{7}$$

where C is a positive value used for regularization. For simplicity, we only present the linear case but to deal with non-linearities one can resort to a kernelized support vector machine (SVM). It is noted that SVM method is also used for classification problem (Cortes & Vapnik 1995).

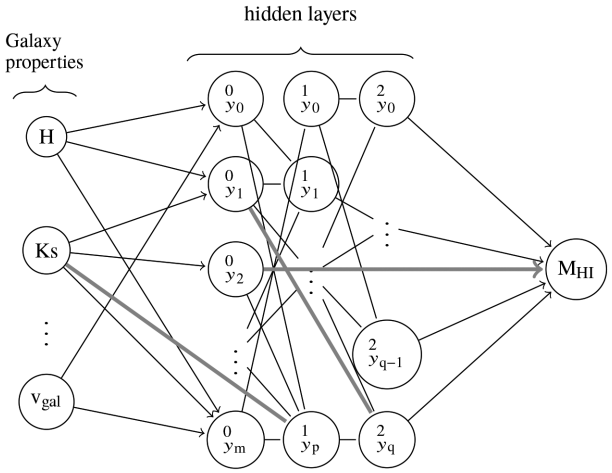

Figure 1. Network graph of our four-layer perceptron with 1 output unit. The hidden layers contain m, p, q neurons, respectively.

4.5 Artificial neural network

We dedicate this section for a rather extended description of the deep neural network (DNN) used for this work. This is so due to its novel application in astronomy. This is not so much the case with other ML techniques described before, as they are at some point fully or partly used to analyse astronomical data.

Due to our hardly correlated features and target, the choice of model to learn the connection between them is very complex, though our maximum number of galaxy properties are limited to only 12 components. Fig. 1 shows a summary of our multi-layer perceptron model. The left nodes show our galaxies properties as input into our three hidden layers and the right most node is the output. ${}^{j}y_k$ represents the kth neuron in the jth layer and is the linear weighted sum of the preceding neurons as shown in equation (8), f_a being the activation function (see Section 4.5.2):

$${}^{j}y_k = f_a\left(\sum_l w_{k,l}^j \times {}^{j-1}y_l + b_k^j\right), \tag{8}$$

where $w_{k,l}^j$ and b_k^j are the weight and bias of ${}^{j-1}y_l$ on ${}^{j}y_l$.

A DNN is then to learn the (close to the) correct values of w's and b's for the model to be able to reproduce the *target* given the *features*.

The choices for the number of the hidden layers, the activation functions between layers and the optimizer are described in the following subsections.

4.5.1 Hidden layers

One of the toughest step that one has to overcome in building a DNN model is the choice of the number of hidden layers and the respective number of neurons in each layer. The use of models with a single hidden layer or the so-called *universal approximators* has been advocated since the artificial neural network was used into solving physical problems. Cybenko (1989) stated that a single hidden layer in a feedforward[9] neural network is enough to capture the continuous non-linearity between the inputs and the outputs. This conclusion was extended later on by Hornik (1991) that the nature of the feedforward structure drives its universality irrespective to the activation function as long as the latter is continuous, bounded, and non-constant (see Section 4.5.2). The 'universal approximation' principle ended recently after the work done by Hinton, Osindero & Teh (2006). They explored the improvement of the multihidden-layer architecture and concluded the following. Although a single hidden layer with finite number of neurons can be enough to map the connection between the input(s) and the output(s), one extra layer is useful to increase the accuracy of the mapping. Any additional layer is only for the model to explore possible representations of the map and to decrease the learning time given a set of data.

For those reasons, and after a trial-and-error approach, we opt to use three hidden layers in our model. We use 100 neurons in each layer to correctly map the galaxy properties with all their possible combinations. We have extra nodes to account for some degrees of freedom for safety. Using different (simpler) configurations end up with similar results for some of the proposed set-ups (zSCmb for instance, see Table 1), but our choice for more complex network is driven by the need for more stable algorithm.

4.5.2 Activation function

Given a set of values fed to one node in our model (see Fig. 1), one has to decide how much of that information should be passed to the next connected node(s). This can be defined with an activation function. A sigmoid function was widely used in the past. Problems occur with that function when the input values of a node are high (or small in the negative end): that is, the vanishingly small gradient at those ends. In our model, we use a rectified linear unit function (RelU, see equation 9). It means that any negative values passing the nodes are set to zero (ignored):

$$f(x) = max(0, x). \tag{9}$$

We also tested the use of an exponential linear unit function (elU, see equation 10). In this case, we allow a small fraction of the negative signal to go through the next connected node(s):

$$f(x) = \begin{cases} x, & \text{if } x \geq 0. \\ \exp(x) - 1, & \text{otherwise.} \end{cases} \tag{10}$$

[9] Any connection between neurons does not form a cycle.

Our test did not get any improvement (if not deterioration) in using eLU. Using different activation functions such as hyperbolic tangent, gaussian, or multiquadratics are not favoured in our case.

4.5.3 Optimization

After each step of calculations, the network should optimize the model based on its current and previous states to improve the subsequent mapping. Our model utilizes a computationally memory efficient optimization due to its dependence to only the first-order gradients, namely the 'adaptive moment estimation' (or Adam). For more details, we refer the readers to Kingma & Ba (2014). Adam optimization, compared to other gradient-based optimization, is very suitable for noisy and sparse gradients, and for simulated data that show very large scatter with respect to a given quantity of parameter (Kingma & Ba 2014). With this optimizer, we have to decide few parameters in advance. The learning step α and the parameters controlling the moving averages of the first- and second-order moments, namely β_1 and β_2 (both $\in[0,1)$), respectively. For this purpose, we chose to minimize the MSE between the target and the prediction from the model: in what follows, we will alternatively call the MSE the 'objective function' $f(\mathbf{x})$: with $\mathbf{x}$ the parameters of the model to be updated, such as weights and biases. At a given time $t \leq T$, where T is the maximal learning time-step, we can update the parameters of the model as shown in the following:

$$g_t = \nabla_x f(x_{t-1}), \tag{11}$$

$$\mu_{1,t} = \beta_1 \times \mu_{1,t-1} + (1 - \beta_1) \times g_t, \tag{12}$$

$$\bar{\mu}_{1,t} = \mu_{1,t}/(1 - \beta_1^t), \tag{13}$$

$$\mu_{2,t} = \beta_2 \times \mu_{2,t-1} + (1 - \beta_2) \times g_t^2, \tag{14}$$

$$\bar{\mu}_{2,t} = \mu_{2,t}/(1 - \beta_2^t), \tag{15}$$

$$x_t = x_{t-1} - \alpha_t \times \bar{\mu}_{1,t}/(\sqrt{\bar{\mu}_{2,t}} + \epsilon), \tag{16}$$

where $\alpha_t = \alpha\sqrt{1 - \beta_2^t}/(1 - \beta_1^t)$ is the time-step at t. Equation (11) shows the gradients of the objective function at t with respect to the model parameters. Equations (12) and (14) update the estimations of the first and second moments. Our moments are biased towards the initial values; thus, we require equations (13) and (15) to account for the corrections. Finally, we update the model parameters with equation (16).

We do not claim that the choice of parameters implemented in our models as well as their configurations are the best to do similar work. We will likely continue to improve this method in subsequent papers.

The reasons we opted for such diversity of regressors in this study are as follows:

(i) To explore the linearity in the data by using LR and SVM
(ii) To explore the power of the ensemble learning with RF and GRAD
(iii) To explore the simplicity, versatility, and speed of k-NN
(iv) To explore the power of the more sophisticated DNN

5 H I PREDICTION USING ML

Our goal is to predict the H I richness of a given galaxy based on its optical/near-IR photometry. We choose to predict H I richness and not H I mass, as it is expected to correlate more with galaxy colours,

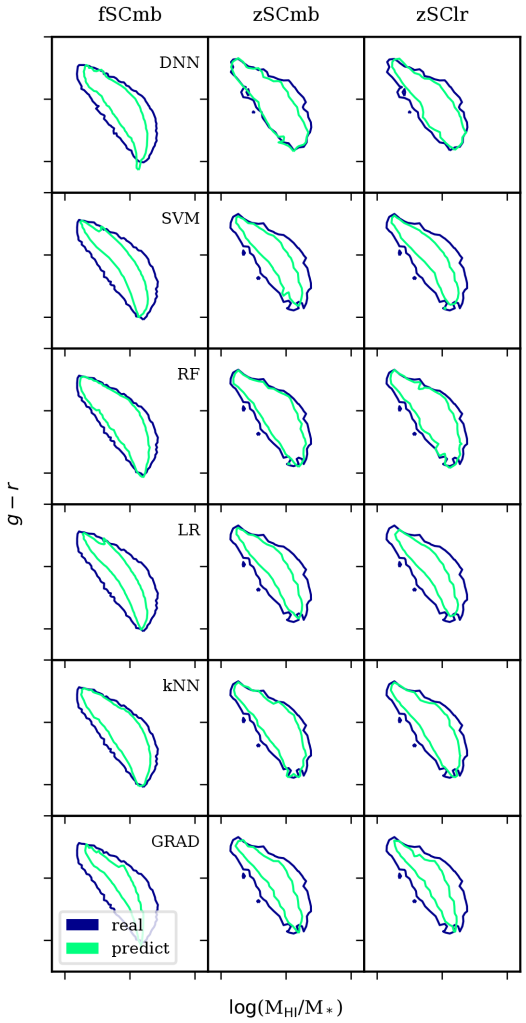

Figure 2. Superposition of the predicted (green) and the real (blue) H I richness of our galaxies (*x*-axes) versus $g - r$ colour (*y*-axes). The contours are enclosing 2σ of the distributions. Each row shows different mapping corresponding a particular method and each column a different set-up (see Table 1).

with H I-poor galaxies being redder than H I-rich ones, so in some sense gives more physical information than just H I mass alone that approximately correlates with stellar mass. None the less, our approach could equivalently be used for either, and we have tested that the resulting accuracy of the predictions is similar.

5.1 Quantifying the mapping accuracy

For a given trained model (see Section 3), we can predict the H I richness of a test set that contains the feature parameters, similar to those used during the training, and the real H I richness. One can then see for a given example (composed by the features) how the model estimates the corresponding H I richness and compare the predicted value of this latter with its real value. Fig. 2 shows the galaxies' $M_{\rm H\,I}/M_*$versus a selected colour $g - r$, one of our input features. The simulated targets are shown with the blue contours and the predicted values with the green contours. Each column represents three selected set-ups (see Table 1) that only use SDSS magnitudes during the training, whereas each row corresponds to

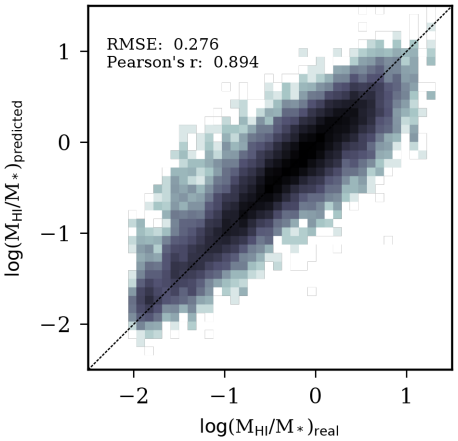

Figure 3. Two-dimensional distribution of the real (*x*-axis) versus predicted (*y*-axis) H I richness with the $z = 0$-trained DNN model, using the zSCmb training set.

one training model. The *z*-trained models shown here (two right columns: zSCmb, zSClr) are at $z = 0$.

Overall, the ML-predicted values follow the true values from the simulation, and show that galaxy colour is anti-correlated with $M_{\rm H\,I}/M_*$ as expected. The mean trend is always well recovered using any of the predictors. However, the scatter in the data is not fully captured by any of the models. The green contours are always inside the blue contours. Different ML algorithms perform differently in this regard. We see that for DNN, RF, and *k*-NN, the two contours are quite close. Only looking at the *f*-trained models (left column), where we train on all the data from $z = 0$–2 simultaneously, it is evident visually that RF maps $g - r$ best, *k*-NN comes next followed by DNN. For the *z*-trained models, where we train individually at various redshifts, DNN, RF, and *k*-NN do similarly well with zSCmb but the performance of RF is better with zSClr (where we add in the colour indices). In contrast, SVM, LR, and GRAD have difficulty to capture the scatter in the data; hence, their predictions tend to be more tightly confined around the mean. While we have shown this specifically for $g - r$, the results for other colours are similar, and typically show that RF and DNN perform the best, with *k*-NN not far behind.

Fig. 3 shows a direct comparison between the real and the predicted H I richness of the galaxies with the DNN models trained and tested with $z = 0$ simulated data. The dashed line shows the 1:1 line; if the ML algorithm were perfect, all points would lie along this line. The correlation is apparent and generally follows the identity line, indicating that the training performs reasonably well in the mean. However, there is a significant scatter, which degrades the performance on a galaxy-by-galaxy basis. The best-fitting slope is also not identically unity, so the correlation is not perfect even in the mean. We thus would like to quantify our regressors' performance using the slope and tightness of the correlation.

To quantify the performance of our ML framework, we choose three metrics:

(i) The slope of the linear mapping $f : y \rightarrow \hat{y}$, where an ideal mapping would have a unity slope.

(ii) RMSE (root mean squared error), given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where y and $\hat{y}$ are the real value and the estimate, respectively, gives the average difference between the predicted and the real values. The square of this metric is also used as a cost function to be minimized in some methods for regression (e.g. DNN and LR). The lower the RMSE, the better the performance of the model is.

(iii) Pearson product–moment correlation coefficient (Pearson's **r**) that tells how scattered the predictions are compared to the true values. The closer to 1, the tighter (or better) the prediction is:

$$\text{Pearson's } \mathbf{r} = \frac{\sum_{i=1}^{N}(y_i - Y)(\hat{y}_i - \hat{Y})}{\sqrt{\sum_{i=1}^{N}(y_i - Y)^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \hat{Y})^2}},$$

where Y and $\hat{Y}$ are the mean values of y_i and $\hat{y}_i$, respectively.

In Fig. 3, we get RMSE=0.276 and Pearson's $\mathbf{r} = 0.894$ for the particular choice of the DNN regressor and the zSCmb training set; this is one of our best cases, but RF is actually slightly better. Previous work by Zhang et al. (2009), estimating H I-to-stellar mass ratio using analytic equation leads to 1σ scatter >0.3, which shows that our ML approach is more accurate.

Fig. 4 shows the performance of the various models considering each set-up in Table 1, using RMSE and Pearson's **r** coefficient. The two columns from the left are the RMSE, and the two columns from the right are Pearson's **r**. Each row corresponds to the results from different features used in the training. The name of the set-up is shown on the top left of each panel. Different results from different learning techniques are presented with the colour coded lines (with distinctive markers). In the following subsections, we discuss how well our various regressors perform when varying the training set and the training method.

5.2 Dependence on redshift

Examining the leftmost column in Fig. 4, these are the RMSEs for various ML algorithms when training on the entire data set from $z =$ 0–2 without any redshift information (*f*-training). The results bear out the trends noted in Fig. 2. The RF method generally does the best (lowest RMSE) for any of the input data sets, while DNN and *k*-NN follow, and then the remaining methods. The RF values are still typically above 0.3, with the lowest values for the fSCmb (SDSS colours, magnitudes, and environment) and perhaps marginal improvement in fAMg that adds the near-IR photometry.

The third column shows the corresponding Pearson's **r** values. The basic story is the same that RF provides the best prediction, with values of $\mathbf{r} \approx 0.85$ in the best cases, with others down to $\mathbf{r} \approx 0.75$. The predictions from the aggregate data set clearly contain significant information, but are perhaps not as optimal as one might get from including some redshift information.

The second and fourth columns show the result of training and testing at individual redshifts (z-training). It is clear that from $z \sim$ 0–0.5, the z-training performs better than the aggregate (*f*) training, with lower RMSE around 0.25 in the best-case RF models (zSCmb and zAMg). The other ML algorithms are clearly poorer than RF, although DNN does reasonably well in the zSCmb case. Similarly, the fourth column showing the Pearson's **r** also is very good at $z =$ 0–0.5, and here, DNN in many cases does nearly as well as RF.

Beyond $z > 0.5$, all the regressors show degrading performance, with increasing RMSE and decreasing **r**. This increase in RMSE likely owes to the fact that at high z, all galaxies are more H I rich ($M_{\rm H\,I}/M_* > 10^{-2}$; Rafieferantsoa et al. 2015), with fewer and fewer quenched galaxies with very low $M_{\rm H\,I}/M_*$. Because the intrinsic $M_{\rm H\,I}/M_*$ versus mass (and other properties) thus becomes fairly flat, it becomes increasingly difficult for the ML to pick out the correct $M_{\rm H\,I}/M_*$ based on other galaxy properties as would be reflected in the photometry. This is likely an intrinsic limitation of this method, owing to the evolution of H I in galaxies.

Redshift information can be obtained observationally, amongst other methods, from photometry or spectroscopy. The latter is still easier to retrieve than direct H I data, while the former typically obtains redshift errors of a few per cent, which is still good enough to ascribe a training redshift. It is clear from the above results that redshift information is useful to improve the predictions. Even out to $z \sim 1$, the limit of currently planned surveys, the predictions do not degrade greatly, it is only at $z > 1$ that they become worse than the aggregate case. Hence from here on, we will primarily discuss the z-training results.

5.3 Dependence on input features

The different rows in Fig. 4 show the impact of varying the input features into the ML framework. As we have seen, RF generally performs the best followed by DNN. GRAD, *k*-NN, LR, and SVM perform similarly poorly regardless of our set-ups (their RMSEs $\simeq 0.34$), with perhaps GRAD performing the worst. For this reason, unless otherwise stated, we are only going to discuss RF and DNN in what follows.

At $z = 0$, using only SDSS magnitudes results in relatively poor performance, with RMSE≈ 0.3 for RF and 0.35 for DNN and others. For RF, using either `color indices` instead of magnitudes (zSCls) or in addition to magnitudes (zSCmb), or including additional magnitudes into the near-IR (zAMg) improves this significantly, with RMSE as low as 0.25 and $\mathbf{r} > 0.9$. Thus, it appears that providing colour information directly into the ML algorithm helps it determine a better mapping than only providing the magnitudes, even though in principle the magnitudes contain all the colour information. Also, providing additional near-IR bands seems to be advantageous.

For DNN, the story is slightly different. Again, only SDSS bands has the worst performance, but here, including the near-IR data does not improve things as much as providing `color indices`, and particularly providing both `color indices` and magnitudes together (zSCmb), which achieves a performance approaching that of RF.

The redshift dependence of RMSE and **r** is similar amongst all these combinations of input data sets. The overall message is that providing more bands is better, which is unsurprising, but also that it is preferable to provide the colours directly rather than the magnitudes given the choice. In many cases, it is possible via SED fitting to obtain a galaxy colour that has uncertainties that are smaller than would be obtained by just subtracting magnitudes, so this may be a more valuable input for ML predictions.

5.4 The slope of the mean relation

In Fig. 5, we show linear fittings for the correlation between real (x-axis) and predicted (y-axis) values for $M_{\rm H\,I}/M_*$. The top panels are for the z-trained models at $z = 0$ and the lower panels for *f*-trained models. Each column corresponds to a given regressor as labelled on top. In each panel, the dark (light) lines represent the 1σ (2σ) contours between the targets and the predictions. The numbers on the top right are the slopes of the linear fits (colour coded) for the two contours. The thick dashed line shows the one-to-one relation, which would be the perfect prediction. We only show the SDSS combined set-up (zSCmb) here, i.e. the features are

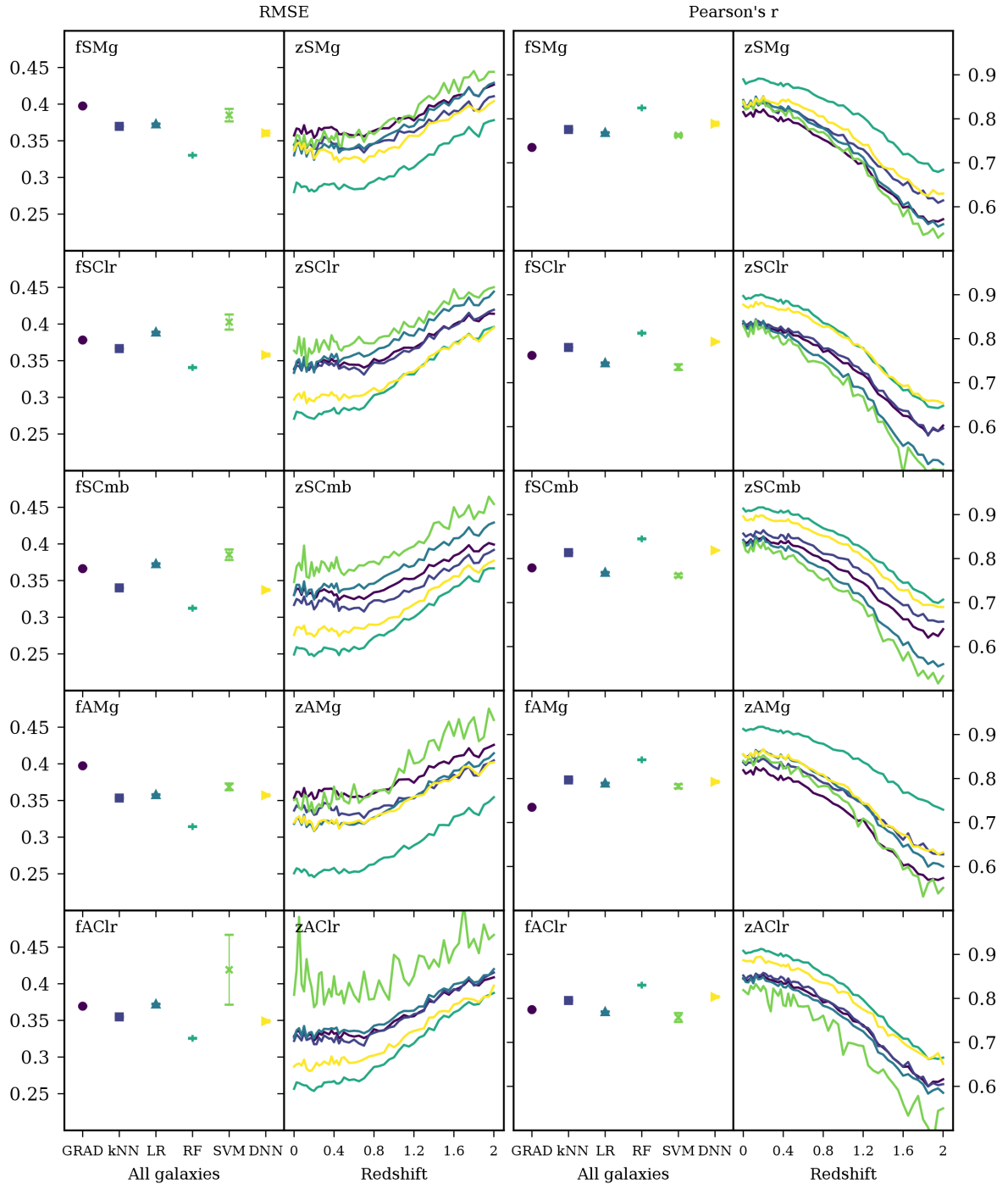

Figure 4. RMSEs and Pearson product–moment correlation coefficients **r** are shown on the two columns from the left and right, respectively. Models perform better if they show lower RMSE and higher **r**. The first on the left shows a mapping for all the galaxies, and the second for galaxies at different redshifts. The dots and lines are colour coded by the training models we use. Each rows show different results for different set-ups. The RMSE values are shown on the left *y*-axes and the **r** values on the right *y*-axes.

SDSS magnitudes+`color indices` $+v_{\rm gal}+\Sigma_3$, but the results from other set-ups are similar.

We can see that *f*-trained (lower panels) models tend to have slopes further away from unity compared to those from the *z*-trained ones. This confirms what we found previously with RMSE and Pearson's **r**, that at low redshifts, training on the smaller but more homogeneous sample at a given z provides a better prediction than training on a larger sample that conflates all the redshifts.

Amongst regressors, again we see that RF and DNN have slopes that are closest to unity, and thus perform better. All other methods have best-fitting slopes below 0.8. However, all the slopes are <1, which indicates an underprediction of the H I richness for H I-rich galaxies and overprediction for H I-poor galaxies. This reflects the fact that, as seen in Fig. 2, the true scatter in the $M_{\rm H\,I}/M_*$ around the mean is not fully reproduced in the predictions, such that all the regressors tend to fit galaxies closer to the mean. Hence, at the lowest

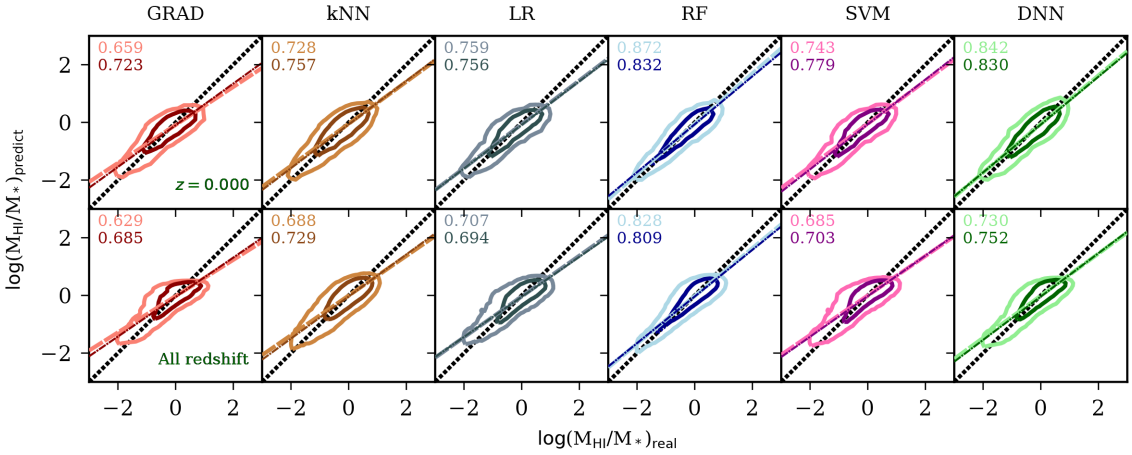

Figure 5. Two-dimensional representations of our real (*x*-axes) versus predicted (*y*-axes) values of H I richness. Upper panels show for different models at $z = 0.0$, whereas the lower panels show for all redshift combined. We only show the results from our {*f*,*z*}SCmb features. The numbers with dark (light) colours on the left top corners show the slopes of the linear fit of the 1σ (2σ) subsample.

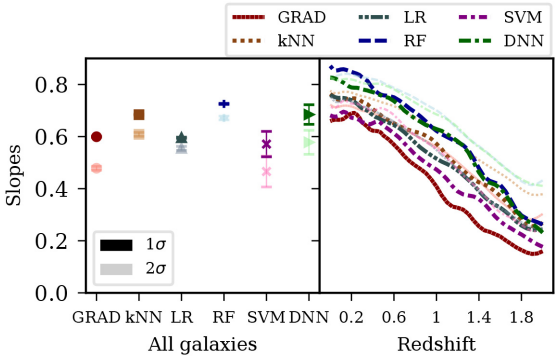

Figure 6. Slopes of the linear fit (*y*-axes) of the relationship between the predictions and the real H I richness of our simulated galaxies. The dark colour (or thick lines) shows the fit for the 1σ sample around the maximum and the light colour (or thin lines) for 2σ. The left-hand panel (*x*-axis showing the names of the models) is similar to what is shown in Fig. 5 (second row), and the right-hand (*x*-axis showing the redshift values) panel presents the evolution of slopes from our zSCmb features.

$M_{\rm H\,I}/M_*$, they tend to fit slightly higher values, while at the highest $M_{\rm H\,I}/M_*$, they tend to fit slightly lower values, resulting in a subunity slope: akin to an Eddington bias. The slope thus partly reflects a measure of how well the scatter around the mean is predicted. The fact that RF and DNN have the best slopes just quantifies the qualitative impression from Fig. 2 that these regressors reproduce the extent of the scatter most closely.

Fig. 6 shows the comparison of slope values for the *f*-trained sample (left-hand panel) and the redshift evolution of the *z*-trained sample (right-hand panel) amongst the various regressors. The left-hand panel effectively just shows a plot depicting the numbers in the bottom row of Fig. 5. Here, RF performs the best but not so far from DNN (considering the variance amongst 10 subsamples), and the other models perform somewhat worse.

The right-hand panel extend the values shown in the upper panel of Fig. 5 to higher redshift. Dark colours (or/and thick lines) show the 1σ slopes and the light colours (or/and thinner lines) show the 2σ slopes. Looking at the *z*-training results (right-hand panel), it is very clear that the slopes of RF and DNN are closer to unity than the other models, and that is true across all redshifts. The 2σ slopes (light colour lines) are generally better than the 1σ's, except at the lowest redshifts. Slopes <0.5 implies a weak correlation between the predicted and the real values of H I richness, so Fig. 6 indicates that all regressors become unreliable beyond $z \gtrsim 1$.

In summary, *k*-NN, RF, and DNN methods show better performance as compared to SVM, GRAD, and LR (Fig. 2). DNN and RF tend to perform better when providing galaxy colours as opposed to photometry, and when providing more bands. Amongst our tests, the best mapping of H I richness was achieved with RF at $z = 0$ using optical and near-IR bands, which gave RMSES ≈ 0.25 and $\mathbf{r} > 0.9$. Using all data from $z = 0$–2 together did not provide as a good fit as training at individual redshifts, despite the smaller samples for the latter. The evolution of RMSE or Pearson's $\mathbf{r}$ shows a stronger redshift dependence beyond $z \sim 0.5$–1 making the prediction uncertain at higher redshift ($z > 1$, see Fig. 4). Slopes of linear fits are generally less than unity owing to the fact that the true scatter is not fully spanned by the prediction; again, RF performs the best with DNN close behind, and the other regressors significantly poorer. All slopes move further from unity with increasing redshift, once again limiting applicability at $z \gtrsim 1$.

6 APPLICATION TO OBSERVED DATA

We now apply and test our ML methodology against real observations from the RESOLVE and the ALFALFA data. RESOLVE survey provides both photometry and $M_{\rm H\,I}/M_*$, so provides an ideal sample to test the efficacy of our predictions. There are two ways we will test this. First, we will train on the RESOLVE data itself, and predict the RESOLVE data, to test how well it works in the ideal circumstance of having the training and testing set be from the same sample. Secondly, we will train on the simulation and predict the RESOLVE data, which is more like the application envisioned for this technique, to see how much degradation there is when the training and testing sets are different. If the simulation was a (statistically) perfect representation of the RESOLVE data, we would expect the resulting RMSE and $\mathbf{r}$ to be similar, but given that we expect some differences, we aim to quantify the degradation in a real-world situation. We concurrently apply the above procedure to the ALFALFA data in order to test sensitivity to different input data sets having different systematics.

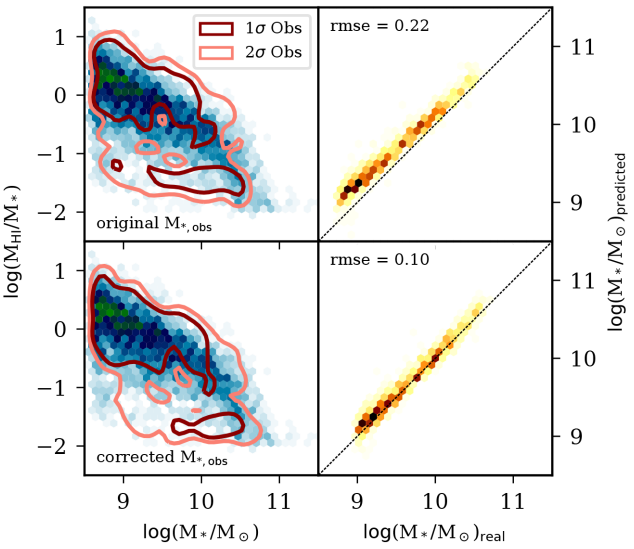

Figure 7. Left-hand panels: The blue-green maps show the distribution of M_* (x-axes) versus$M_{\rm H\,I}/M_*$ (y-axes) of the simulated galaxies, while the dark and light red contours show the 1 & 2σ distributions of the RESOLVE data. Right-hand panels: distributions of the real (x-axes) and predicted (y-axes) galaxy stellar masses of the RESOLVE galaxies. Upper panels show the distributions prior to the correction to observed stellar masses as described in the text, and lower panels after correction. The lack of bimodality in the simulated data (right-hand panels) as seen in the data is mainly due to our cut to only include galaxies with $M_{\rm H\,I}/M_*>10^{-2}$.

6.1 Simulated versus observed data

6.1.1 RESOLVE data

We first describe the RESOLVE data. We make use of the photometry data (Eckert et al. 2015) as well as their corresponding H I flux (Stark et al. 2016) from the Data Release II of the RESOLVE survey. We use the following standard equation:

$$M_{\rm H\,I} = 2.36 \times 10^5 \times D^2 \times F_{\rm Total} \quad (17)$$

to compute the H I mass in $M_\odot$, where D is the distance to the galaxy (Mpc) calculated from the apparent and absolute magnitude in **r** band given in the photometry data. $F_{\rm Total}$, provided by the RESOLVE data, is the total H I line flux ($Jy.\,{\rm km\,s^{-1}}$) of the galaxy. The RESOLVE photometric data release[10] contains SDSS (*u,g,r,i,z*), 2MASS (*J,H,K*), GALEX (*NUV*), and UKIDSS (*Y,H,K*) band magnitudes.

One immediate issue when comparing to simulations will be that stellar population models, initial mass function, etc., used to obtain M_* from the data (from which we compute $M_{\rm H\,I}/M_*$) is different between what we assume in LOSER versus what RESOLVE assumed to obtain their M_* values. Hence, it turns out there is a small offset in M_* that we must first correct. We do so empirically, by using our ML framework to predict the M_* from the photometry in our simulations and from RESOLVE, and then comparing the M_* values.

Fig. 7 (right-hand panels) shows the difference between the original (top) and the corrected (bottom) M_* values from RESOLVE. The original RESOLVE data are offset by ∼0.1–0.2 dex; this is within the uncertainties of typical M_* determinations from photometry. The correction we apply is a linear scaling of the stellar masses to match with MUFASA galaxies, obtained by training the DNN model with the simulation to predict the stellar mass of the RESOLVE data, and comparing the result with the real value from RESOLVE. We repeat the process 10 × and take the average of the linear slopes and the intercepts to obtain the following relation: $\log M_{*,\rm corrected} = 0.920 \times \log M_{*,\rm original} + 0.924$. It can be seen that M_* is predicted very tightly, with a scatter of RMSE =0.1 once the correction is applied. Prior to the correction, the RMSE =0.22 relative to the 1-to-1 line, which is dominated by the offset rather than the scatter itself. Note that scaling the simulated stellar masses would give the same results, but we do not use this option because we know exactly the stellar mass of the simulated galaxies.

We can also compare the trend of $M_{\rm H\,I}/M_*$ versus M_* in the simulations and RESOLVE, which is done in the left-hand panels of Fig. 7, before (top) and after (bottom) the M_* correction. The green-blue distributions on the left-hand panels are from MUFASAgalaxies, whereas the contours are from the observational data. In general, particularly after the correction is applied, the simulations and observations agree quite well for the bulk of the galaxies. A clear trend is seen that lower-M_* galaxies have higher H I fractions. The mean trend of the galaxies with H I is in good agreement between RESOLVE and this simulation, which confirms the agreement versus other data sets shown in Rafieferantsoa et al. (2015). This indicates that MUFASA provides a generally viable model to predict observed H I from photometry.

There is a notable difference that the observational data shows a bimodal distribution that is not seen in the simulated data. This is because we have explicitly ignored galaxies from MUFASA with $M_{\rm H\,I}/M_*<0.01$. In MUFASA, we have many galaxies with no H I, while in the observations there is a distribution of low-$M_{\rm H\,I}/M_*$ values. We will leave more careful modelling of these low-$M_{\rm H\,I}/M_*$ objects for future work, but we note that the bimodality is going to degrade our results since the ML is unlikely to effectively predict galaxies with $M_{\rm H\,I}/M_*$ approaching ∼0.01.

We also check if the range of magnitudes between the RESOLVE and MUFASA are in broad accordance. Fig. 8 shows the same distributions as in Fig. 7 (lower left-hand panel), except that now the colours in each hexagonal bin represent the mean magnitudes of the galaxies in that bin. We show *ugriz* magnitudes for illustration, but we get similar results for other bands. Each column represents one band. Upper and lower panels are for simulated and observational data, respectively. We can clearly see that the trends are consistent. Note that apart from SDSS magnitudes, we also use *NUV, J, H,* and K_s magnitudes in the training. We however point that including all those bands decreases the size of the sample due to missing data in each band. The RESOLVE data contain 2159 galaxies with SDSS magnitudes. When accounting for *NUV, J, H,* and K_s, we end up with only 1017 galaxies.

6.1.2 ALFALFA data

We use the α.100 ALFALFA data (Haynes et al. 2018) that contain the derived H I mass and the position of the sources in RA and Dec. To obtain the photometric magnitudes (u, g, r, i, z) and the stellar masses, we cross-matched the sources to their SDSS counterpart based on the positions. We allow a 6 arcsec maximum search radius. We end up with 16 588 galaxies. We use the pre-built SDSS web interface[11] to do our cross-matching procedure. Similar to RESOLVE data, we need to account for the stellar mass correction

[10] https://resolve.astro.unc.edu/data/resolve_phot_dr1.txt

[11] http://skyserver.sdss.org/dr14/en/tools/crossid/crossid.aspx

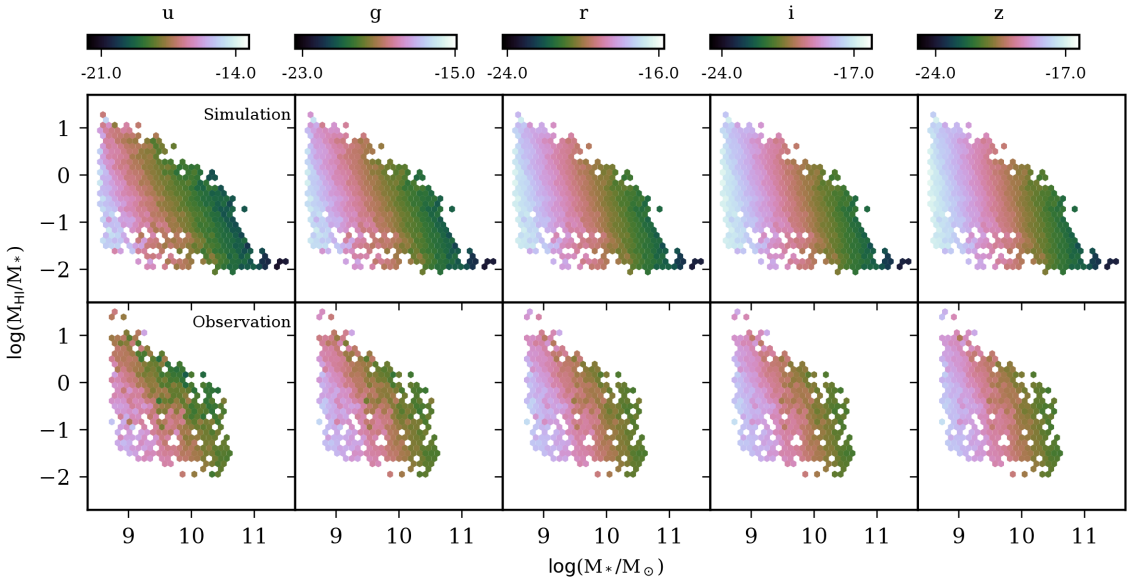

Figure 8. The *x*-axes and *y*-axes represent the stellar masses and H I richness of the galaxies, respectively. Similar to bottom left-hand panel of Fig. 7, but showing the mean magnitudes in each pixel for the SDSS passbands (columns). The top and bottom panels are for simulated and observational data, respectively. The agreement between the two data is noticeable and the range of the observational data are well included in that of the simulated ones.

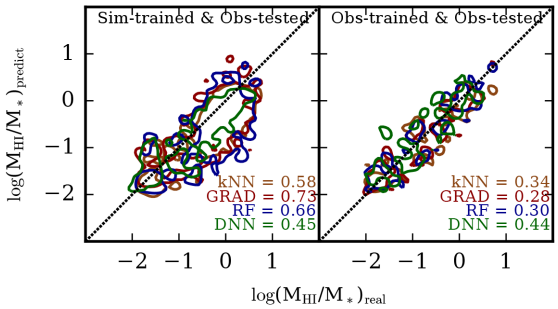

Figure 9. RESOLVE data. Predictions of the observed H I richness (*y*-axes) using different mapping algorithms (colour coded lines). Left-hand panel shows the results when the algorithms are trained with simulated data. Right-hand panel shows the results when training the algorithms with observational data. The contours correspond to 1σ distribution.

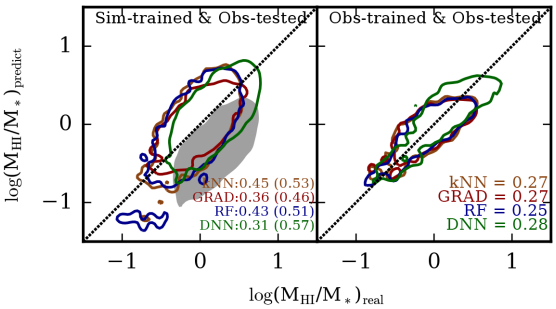

Figure 10. Similar to Fig. 9 but with ALFALFA data instead. On the left-hand panel, the numbers in the brackets are the RMSE prior to the linear shift of the predicted values (see the text). The contours correspond to 1σ distribution. The grey shaded area on the left-hand panel corresponds to the DNN (regressor with the highest offset) prediction prior to the shift. Other regressors' (not shown) have lower offsets.

for SDSS data using the same method described before and obtain $\log M_{*,\mathrm{corrected}} = 1.022 \times \log M_{*,\mathrm{original}} - 0.4534$.

6.2 Training on and predicting observed data

6.2.1 RESOLVE results

We first consider the case where we train the regressors using one subset of the RESOLVE data and test them using the other subset (the one which was not used for the training). Due to the relatively small sample in hand, we only use 10 per cent of the data for testing. This case can be considered optimal in the sense that the training and testing sets are drawn from (different parts of) the same sample, so there are no systematic differences.

The right-hand panel on Fig. 9 shows our prediction using the test sets. Judging by the contours, it is clear that all the presented models here perform reasonably well, i.e. the distribution of the real versus predicted values lie along the identity line, and the predicted values (*y*-axis) cover all the range of the real values for all regressors. Comparing regressors, GRAD with RMSE=0.28 performs best followed closely by RF, *k*-NN, and lastly DNN with RMSE=0.44. Now the trend is reversed such that DNN, which was amongst the best in the previous scenario becomes the worst in this case. DNNs typically require larger training samples to properly constrain the large number of layers, so it is likely its poor performance owes to the small sample of RESOLVE galaxies.

6.2.2 ALFALFA results

The previous results motivate us to test if our models produce similar outputs using a different set of data. Using the larger ALFALFA sample cross-matched with SDSS photometric data, we can equally perform the previous exercise. We train the regressors on the training sample (75 per cent of ALFALFA) and predict the H I richness on the testing sample (remaining 25 per cent).

The right-hand panel of Fig. 10 shows our results. We attain RMSE <0.3 for all regressors shown here, and none show any sys-

tematic bias, with a median lying on the one-to-one line across all H I richness. In detail, RF performs slightly better than others, whereas DNN slightly worse than others. The better performance, compared to the one corresponding to the RESOLVE sample, likely arises in part from the almost eight times larger sample. We note that the difference in the ranges of real H I richness used in DNN and the others is due to the different random splitting between `TensorFlow` and `scikit-learn`, but the resulting performance shows no qualitative difference.

These tests show that our ML framework is able to quantitatively predict observed data when (a distinct subset of) the same data set is used for training. Similar results has been shown in other works such as Teimoorinia et al. (2017), and in fact their predictions are formally even tighter, likely owing to the inclusion of derived galaxy properties in addition to purely photometric data. While these tests are encouraging, our broader aim is to utilize our ML framework to make predictions in regimes where no training set exists, by using our simulated galaxies as the training set. Thus, we must now test the case where we train on the simulations, and test against the observations from RESOLVE and ALFALFA.

6.3 Training on MUFASA and predicting observed data

A more general application would be where we have no or very limited H I training data, and only photometric data. This might be the case at $z \sim 0.3$–1, where the H I data are almost non-existent now and even future surveys will provide only a sparse sampling of the most H I-massive objects. In this case, we would like to be able to use the simulations to provide the training set. Naturally, this introduces more uncertainties and assumptions, because the simulations build in a specific physical model that likely is not exactly correct, and does not reproduce the real H I population in all its details. To test how much more uncertain the predictions would be, we can attempt this using RESOLVE and ALFALFA where we *know* what the correct answer is, and see how well the simulation recovers it relative to the case in the previous section where we used observed data themself to train.

In order to mitigate the effects of those uncertainties, one must carefully mimic the input features of the simulated data to encompass those from the observational data as discussed in the previous section. Given that MUFASA reproduces several observables that are usually used as benchmark for simulation models, such as stellar mass function, H I mass function, and specific star formation rate function (Davé et al. 2016, 2017a; Davé et al. 2017b), we feel confident that it provides a state-of-the-art approach to making predictions for upcoming surveys such as LADUMA or MIGHTEE, i.e. using simulated data for training the algorithms and applying it to available observational photometric data.

6.3.1 Simulation-trained ML applied to RESOLVE

The left-hand panel of Fig. 9 shows the H I-richness prediction of our four best models, training the regressors with the simulation data and predicted the H I richness of the RESOLVE data. The contours show the distributions of the RESOLVE H I richness (*x*-axis) versus the predicted H I richness (*y*-axis) from the models. The numbers on the bottom right of each panel show the RMSE of each model.

Overall, the predictions still lie along the one-to-one relation, indicating that using the simulations to train still provides an adequate prediction in the mean. However, the RMSE values are much higher here than in the right-hand panel. This clearly shows that the simulated sample does not fully mimic the details of the observed sample. Given the discrepancies between simulation and observation, implying differences of the underlying distributions of the two samples, this is not surprising.

k-NN, GRAD, and RF now all have RMSE values above 0.5, which is fairly poor. They estimate with larger scatter and a noticeable offset towards lower H I-richness values, which is strongest at $\log_{10}(M_{\rm H\,I}/M_*) \sim 0$ (lower contours are farther from 1:1 line than the upper ones).

Rather remarkably, DNN (green contour) now performs the best in this case, with RMSE = 0.45 and predictions extending to the lowest values ($-2 \leq$) following the 1:1 line. Although DNN was outperformed in Fig. 4 using only simulated data for training and testing, we can clearly see here that its performance shines in a more difficult scenario, where now the training sample is much larger, but the data are more complex. Indeed, the RMSE for DNN hardly changed at all when using the RESOLVE or MUFASA data to train, though this probably arises from the larger training sample offsetting the less homogeneous testing sample. Our results suggest that in this real-world application, DNN can learn better from the simulated data than simpler regressors. We note that the offset in the predictions will always be present in any pairwise set of different data, be it (simulation observation) or (observation observation).

From those two approaches, left-hand and right-hand panels of Fig. 9, we can see that DNN presents robust predictions regardless of the training set-ups. It is able to learn important features from the simulation and translate those into the observed data. kNN, GRAD, and surprisingly RF are less efficient in doing so. The latter only performs best when the training and testing samples are drawn from the same main sample.

6.3.2 Simulation-trained ML applied to ALFALFA

The left-hand panel of Fig. 10 shows the results when the training is done with MUFASA, and the trained models predict the H I content of the ALFALFA sources. The contours show the distributions of the ALFALFA H I richness (*x*-axis) versus the predicted H I richness from the models. The numbers inside the brackets are the original RMSE before the linear shift (colour coded). The grey shaded area corresponds to the DNN prediction before the shift is applied. Other regressors (not shown for clarity) have lower offsets. The linear shift is the necessary amount of H I richness (in dex) to minimize RMSE: i.e. the required *intercept* value to make the linear fit of the prediction coincide to the 1:1 line the closest. Such offset is expected due to the nature of the ALFALFA survey that was only targeting H I fluxes resulting in a H I biased-high sample. We did not account for such H I offset with the predictions for RESOLVE data due to the latter being highly complete and volume-limited survey data. The H I offsets are highest for DNN (=0.43 dex), whereas the other predictions requires only ∼0.23 dex. We interpret the difference in those values to be a result of the different slopes of the linear fits between the prediction and the original values. They are 1.06, 1.13, 1.15, and 1.13 for DNN, RF, GRAD, and *k*-NN, respectively. Not accounting for the offsets, one might read the best performance to be that of GRAD with RMSE =0.46, followed by RF and *k*-NN and finally DNN with RMSE =0.57. We note that previous work by Rafieferantsoa et al. (2015) showed an offset of >0.5 dex in H I richness between $\alpha.40$ ALFALFA data cross-matched with SDSS and our previous galaxy formation model. In order to account for this properly, we would have to create mock ALFALFA selection functions and 'observe our simulations appropriately, which is a

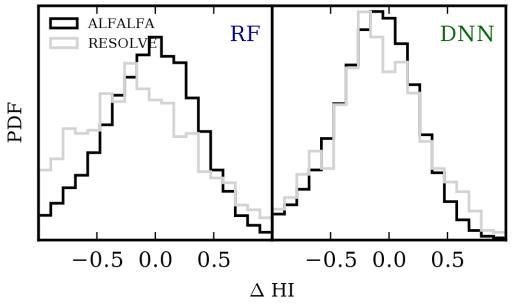

Figure 11. Log-scaled probability distribution function of the uncertainty of the H I predictions from simulation-trained regressors. Left: random forest. Right: deep neural networks. The black (grey) lines are the prediction uncertainties of the ALFALFA (RESOLVE) data.

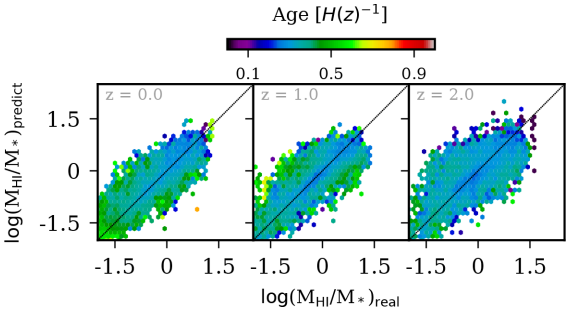

Figure 12. Mean galaxy age for each pixel in the distribution of the real (*x*-axes) and predicted (*y*-axes) H I richness of the simulated galaxies. This result is from the DNN-trained model. Different panels show for different redshift. We use the age of the galaxies at the given redshifts (shown on the top left corner in each panel).

challenging task that is beyond the scope of this work, and is not really even feasible given the small simulation volume compared to SDSS.

Fig. 10 shows that the predictions are remarkably good despite training and testing on different data sets. The RMSE $\sim$ 0.31 is minimized again for DNN, and other models perform less well but still have RMSE<0.5. All distributions have medians that closely follow the identity line, showing no systematic offset. This accuracy is significantly better than that obtained using the RESOLVE data. It is unclear what aspects of RESOLVE versus ALFALFA make the latter more closely align with the simulation predictions, since they encompass a similar redshift and mass ranges.

Finally, Fig. 11 shows the probability distribution functions of the deviation of the predictions from the true values for RESOLVE (grey line) and ALFALFA (black line) data. The regressors are trained from MUFASA data. The left-hand panel is for RF and the right-hand panel for DNN. We log-scale the *y*-axis to highlight the peculiarities between the distributions. For both DNN and RF, we notice that the peaks of the pdf are below the mean (ΔH I<0) for the predicted H I values of RESOLVE data. That is already noted in Fig. 9 but it is shown more quantitatively here. For ALFALFA, the peak is closer to the mean ($\sim$0). Predictions of RESOLVE H I also have more extended tails probably due to the size of the data. The higher counts for underpredicted H I content from the RF regressor (RESOLVE, grey line at ΔH I<0) is again already seen in Fig. 9 (bent shape of the contour).

To summarize, we have shown that training on a subset of observational data can yield a reasonably tight prediction for a testing set taken from the same data. This provides a way to populate photometric surveys in scenarios where a sizeable H I training set is available, such as RESOLVE or ALFALFA; similar results have been shown in previous works (e.g. Teimoorinia et al. 2017). The new aspect in this paper involves training the models using simulated data and predicting the observational targets. This can result in somewhat higher uncertainties, but still without any significant systematic offsets. In particular, testing a simulation-trained network on over 16 000 ALFALFA galaxies cross-matched to SDSS shows that the simulation-trained ML produces a scatter comparable to that obtained by training on the ALFALFA data itself, around 0.3 dex, when using DNN. In terms of regressors, while RF has been the best choice in more homogeneous training/test set situations (simulation-simulation and observation-observation), when applying the simulation training to observational data, the performance of our DNN clearly outshines the others.

7 DISCUSSION

Extraction of information in a given set of data is a challenge in all models. Although RF and DNN are our best regressors, they still have difficulty in extracting all the necessary information. That being said, attaining an accuracy of $\mathbf{r} > 85$ per cent is a non-trivial success for both of regressors. In our training for the DNN, we make sure that the loss function stays unchanged for several training steps to make sure the network learns as much information as it needs but not as much as it might overfit the training data and lose the important information necessary for the prediction. It may be possible to tune this better.

It is possible that photometric surveys can yield other information such as the age, star formation rate, and (from a group catalogue) halo masses, albeit with some uncertainties. It is interesting to ask whether providing such information would improve predictions. However, we find that this is not obviously the case. We illustrate this for the mean stellar age in Fig. 12. Here, we show the distribution of the galaxies based on their real H I richness and the predicted values from the DNN model, with the colour of each hexagonal bin showing the mean age of galaxies falling in that bin (in unit of the Hubble time at the given redshift). Different panel shows different redshifts: *left, centre, right* for $z = \{0, 1, 2\}$, respectively. We can see that for a given H I richness value we cannot see any age gradient in the predicted values, and it remains the case up to $z = 2$. We interpret this to mean the ML model has learned about the age of the galaxies even though that information was not explicitly given in the training set. The same situation happens with the specific star formation rates and the halo mass of the galaxies. This is the case for all of our ML models. Hence, providing such information, which introduces further uncertainties from their estimation, is unlikely to be helpful.

Then we might ask why do some regressors perform better than others? We believe that the design of the models themselves may lead to different mapping of the input–output, thus, to improved results depending on the data. Changing the layer structures in DNN or optimizing the tree size (or the number of base estimators) in RF might alleviate certain issues we encountered in our training. We are currently analysing such possibility and might improve our model in that direction in upcoming work. Also DNN may particularly benefit from a larger simulation training sample with more dynamic range than available in MUFASA.

One useful feature of RF is that it provides an estimate of the importance level of the input parameters, based on the rate of incidence

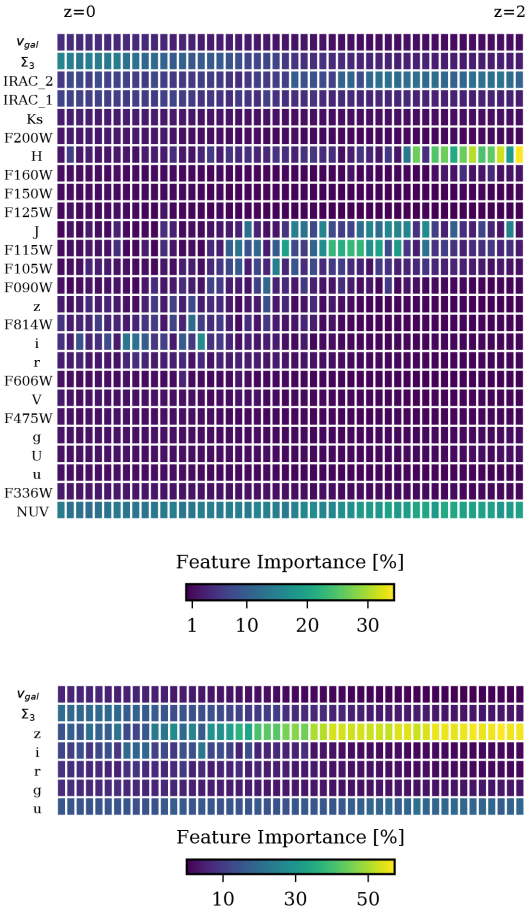

Figure 13. Evolution of the importance of the input features from the RF training. Each row represents one band with the filter name on the left, except for the first (second) row that shows for the line-of-sight velocity (third nearest neighbour) feature. The bands from the bottom to the top are with increasing peak wavelengths. Left to the right shows the feature importance from $z = 0$ to $z = 2$.

that a given parameter is utilized in the decision trees. We show in Fig. 13 the importance of parameters from RF training. The upper subfigure shows the result when using all the available magnitudes from our simulation, whereas the lower sub-figure represents the result when only using the SDSS magnitudes. The first (second) row in each subfigure shows the importance of the line-of-sight velocity $v_{\rm gal}$ (third nearest neighbour Σ_3) from $z = 0$ (left) to $z = 2$ (right). The remaining rows show for bandpass filters (names on the left) with a wide range of peak wavelengths from 2309Å (bottom row) increasing to 44630Å.

It is interesting to see that Σ_3 becomes increasingly important only at later epochs; this is physically expected since environment becomes an increasing determinant of H I properties at lower redshifts (Rafieferantsoa & Davé 2018). The line-of-sight peculiar velocities $v_{\rm gal}$ do not add value to the training, which is unsurprising since it is not obvious why the H I content should care about peculiar velocity (except perhaps through correlations of peculiar velocities and the large-scale potential well); this in a sense serves as a sanity check that our method is not finding physically implausible relationships. In the upper subfigure, the IRAC channels have some importance at higher redshift, particular IRAC 4.5 μm while 3.6 μm is less important. The *H*-band magnitude is very important at high redshift but contributes much less at low redshift. The importance of magnitudes between *i* (6250 Å) and *J* (12 500 Å) bands move from low to higher peak wavelengths towards higher redshift. *NUV* magnitudes seem to exhibit relatively high importance at all redshift bins, highlighting the connection between H I and the gas that fuels star formation and hence UV light.

In the lower subfigure with a more restricted input set, *z* magnitude is very important at higher redshift but becomes less although still important at $z = 0$, whereas the importance of *i* magnitude increases towards the present day. The value that *u* magnitude adds to the accuracy of the prediction seems to be relatively constant at all redshifts, following *NUV* in the upper subfigure. It appears overall that the reddest available photometric band has the highest RF importance level at high redshifts but decreases in importance at lower redshifts, while Σ_3 increase in importance particularly at $z \lesssim 0.5$.

On the whole what the two panels in Fig. 13 tell us is that given the features available in the data, the feature importance in principle allows one to select only a set of the most important ones in order to achieve a given accuracy. This, amongst other methods like Principal Component Analysis (PCA), is of a great value especially when reducing the dimensionality that might not be avoidable due to a limited computing power or when the dimension is as big as the size of the data (i.e. number of features is as large as the number of examples for the training). Also, the importance levels could be helpful in survey design, if a particular photometric band is more useful it might be regarded as higher priority to obtain. However, one must be aware that in many cases, RF importance levels do not truly reflect the necessity of a given data, in the sense that sometimes RF says a particular input is important, but the information from that input is actually encoded in the other inputs, so that removing it does not have as detrimental effect as one might think (e.g. Agarwal, Davé & Bassett 2018). Likewise, one should not expect the importance of input parameters to be continuous across cosmic time because the algorithm can swap between parameters with similar importance. Properly assessing the importance level would involve re-training the entire data set removing each input, in turn, to assess the increase in RMSE, which is highly computationally intensive. None the less, RF importance level can at least provide a guide for this process.

8 CONCLUSION

We have investigated estimating the H I richness of galaxies based on their optical and near-IR survey properties, in particular SDSS $\{u, g, r, i, z\}$, Johnson $\{U, V\}$, and 2MASS $\{J, H, K_s\}$, line-of-sight velocities, and environmental measures, using ML. For our analysis, the training data have been generated from the MUFASA simulation, which has been shown to provide a good description of the H I content of observed galaxies. We have tested various ML regressors including RFs and DNNs. We considered various input feature combinations, including only SDSS magnitudes and environmental properties, using galaxy colours instead of and in addition to magnitudes, and including 2MASS and Johnson magnitudes. We trained each model to predict $M_{\rm H\,I}/M_*$ based on an aggregate of all simulated galaxies at $z = 0$–2 (*f*-training), and in 50 individual redshift bins (*z*-training). As an example application, we applied this framework to the RESOLVE and ALFALFA+SDSS galaxy survey catalogues with H I and photometric data. To measure and compare the performance of each method, we used RMSE, Pearson correlation coefficient **r**, and the correlation slope.

We summarize our main findings as follows:

(i) By using 75 per cent of the MUFASA data for training and testing on the remaining quarter, we find that all ML methods are able to approximately recover $M_{\rm H\,I}/M_*$ from galaxy photometry. The accuracy depends both on the input data set and on the ML algorithm. Generally, RFs provide the best performance at $z = 0$, i.e. lowest RMSE ≈ 0.25, highest $\mathbf{r} \approx 0.9$, and slope closest to unity, with DNN close behind.

(ii) At $z \lesssim 1$, it is advantageous to do the ML training at a given redshift rather than aggregating all redshifts. The smaller number of galaxies available for training in the former is outweighed by the conflating of evolutionary trends when aggregating. The RMSE of all ML algorithms increases with redshift, with commensurately lowered **r** and a best-fitting slope diverging from unity, though the effect is mild out to $z \sim 0.5$. Predictions at higher redshifts are more challenging owing to reduced trend in $M_{\rm H\,I}/M_*$ amongst high-z galaxies, since most galaxies at $z \gtrsim 1$ have similar $M_{\rm H\,I}/M_*$ prior to significant populations of quenched galaxies arising.

(iii) Providing more input training data results in better predictive power, unsurprisingly. Using only SDSS data results in RMSE≈ 0.3 for RF at $z = 0$, while either including 2MASS data or training on both colours and magnitudes yields a more optimal RMSE. DNN has in the best case similar performance, but it is more strongly dependent on the selected input features.

(iv) All the regressors tend to underpredict the high H I richness and overpredict the low H I richness, as shown by the slope (<1) of the linear fits between the targets and the predictions. This owes to the regressors being unable to fully capture the scatter in the $M_{\rm H\,I}/M_*$ values at e.g. a given colour, instead tending to push the $M_{\rm H\,I}/M_*$ towards the mean. This raises the value of low $M_{\rm H\,I}/M_*$ objects and lowers it for high $M_{\rm H\,I}/M_*$ objects, resulting in a sub-unity slope. The underprediction of the high H I richness is more severe at high redshift (Fig. 6).

(v) By training our ML framework on a subset of the RESOLVE and ALFALFA+SDSS data and testing it on the remainder, we showed that it is possible to predict $M_{\rm H\,I}/M_*$ with RMSE $\lesssim 0.3$, which is comparable or better than what is obtained with scaling laws. RF again performs amongst the best, though GRAD and k-NN also show slightly better performance. When training on MUFASA and testing on the observed data, we find the best regressor is DNN. The predictions are significantly degraded with RMSE$\approx$0.45 with RESOLVE data, but for the larger ALFALFA+SDSS sample the RMSE $\approx$0.31 which is only slightly worse than the case where we trained on ALFALFA+SDSS data itself. The worse performance in this case likely owes to subtle mismatches between simulation predictions and analysis procedures versus those from the observations. While the scatter is substantial, the median trend remains well-matched, showing that the ML prediction introduces at most mild systematic biases.

We have shown through this study that it is clearly possible to estimate the H I richness of a galaxy by relying only on the information from photometric magnitudes. We considered various magnitudes from different surveys like SDSS, Johnson, and 2MASS in this work, but including other bands is doable. The broadly successful test of training on simulated data and applying to observed data, particularly ALFALFA+SDSS, suggests that the estimation of H I gas at higher redshift using the methods presented here, even with the lack of testing data, is fruitful. With the advent of future surveys such as LADUMA and MIGHTEE, our ML framework constitutes an important new tool to aid studies of neutral hydrogen evolution in galaxies out to intermediate redshifts.

For our analysis, we have only selected galaxies that are observable in H I, with a threshold of $M_{\rm H\,I}/M_* > 10^{-2}$. This raises a key question: 'Would a model still generalize well if one also included the H I-depleted galaxies in the data set for the training?' There are two ways to address this question. First, we can simply add the H I-deficient galaxies in the data set and redo the fitting procedure prescribed in this work, although from the standpoint of observations, predicting the H I richness of a H I-depleted or gas-starved galaxy is not really meaningful. The more elegant approach would be to first use ML to classify galaxies based on their observable features whether they are H I deficient or not, then only estimate its H I richness (based on the same features) in the case it would potentially contain observable H I. Of course, the minimal value of observed H I can be a free parameters in our model but in reality that should depend on the telescope capabilities.

Future work will discuss these solutions, provide more tailored predictions for upcoming surveys, utilize larger training samples that could particularly help improve DNN results, and make this tool available to the community.

ACKNOWLEDGEMENTS

The authors thank S. Hassan, B. Moews, and G. I. G. Józsa for helpful conversations and guidance. MR and RD acknowledge support from the South African Research Chairs Initiative and the South African National Research Foundation. MR acknowledges financial support from the SKA post-graduate bursary programme. SA acknowledges financial support from the SKA. The MUFASA simulations were run on the Pumbaa astrophysics computing cluster hosted at the University of the Western Cape, which was generously funded by UWC's Office of the Deputy Vice Chancellor. MR uses additional computing resources from the Max Planck Computing & Data Facility (http://www.mpcdf.mpg.de).

REFERENCES

Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410
Altman N. S., 1992, j-AMER-STAT, 46, 175
Breiman L., 2001, Mach. Learn., 45, 5
Catinella B. et al., 2010, MNRAS, 403, 683
Catinella B. et al., 2013, MNRAS, 436, 34
Conroy C., Gunn J. E., 2010, Astrophysics Source Code Library, ascl: 1010.043
Cortes C., Vapnik V., 1995, Mach. Learn., 20, 273
Crain R. A. et al., 2009, MNRAS, 399, 1773
Croton D. J. et al., 2006, MNRAS, 367, 864
Cunnama D., Andrianomena S., Cress C. M., Faltenbacher A., Gibson B. K., Theuns T., 2014, MNRAS, 438, 2530
Cybenko G., 1989, Math. Control Signals Syst., 2, 303
Davé R., Katz N., Oppenheimer B. D., Kollmeier J. A., Weinberg D. H., 2013, MNRAS, 434, 2645
Davé R., Rafieferantsoa M. H., Thompson R. J., 2017b, MNRAS, 471, 1671
Davé R., Rafieferantsoa M. H., Thompson R. J., Hopkins P. F., 2017a, MNRAS, 467, 115
Davé R., Thompson R., Hopkins P. F., 2016, MNRAS, 462, 3265
Duffy A. R., Kay S. T., Battye R. A., Booth C. M., Dalla Vecchia C., Schaye J., 2012, MNRAS, 420, 2799
Eckert K. D., Kannappan S. J., Stark D. V., Moffett A. J., Norris M. A., Snyder E. M., Hoversten E. A., 2015, ApJ, 810, 166
Friedman J. H., 2000, Ann. Stat., 29, 1189
Gabor J. M., Davé R., 2015, MNRAS, 447, 374
Giovanelli R. et al., 2005, AJ, 130, 2598
Hahn O., Abel T., 2011, MNRAS, 415, 2101

Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edn. Springer, New York
Haynes M. P. et al., 2018, ApJ, 861, 49
Hinton G. E., Osindero S., Teh Y.-W., 2006, Neural Comput., 18, 1527
Holwerda B. W., Blyth S.-L., Baker A. J., 2012, in Tuffs R. J., Popescu C. C., eds, Proc. IAU Symp. 284, The Spectral Energy Distribution of Galaxies – SED 2011. p. 496
Hopkins P. F., 2015, MNRAS, 450, 53
Hornik K., 1991, Neural Netw., 4, 251
Jones M. G., Papastergis E., Haynes M. P., Giovanelli R., 2016, MNRAS, 457, 4393
Kannappan S. et al., 2011, Am. Astron. Soc. Meeting Abstr. #217, p. 334.14
Kannappan S. J., 2004, ApJ, 611, L89
Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2
Kingma D. P., Ba J., 2014, CoRR, abs/1412.6980
Krumholz M. R., Gnedin N. Y., 2011, ApJ, 729, 36
Mitra S., Davé R., Finlator K., 2015, MNRAS, 452, 1184
Muratov A. L., KeresŽ D., Faucher-Giguère C.-A., Hopkins P. F., Quataert E., Murray N., 2015, MNRAS, 454, 2691
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Planck Collaboration XIII, 2016, A&A, 594, A13
Quilis V., Planelles S., Ricciardelli E., 2017, MNRAS, 469, 80
Rafieferantsoa M., Davé R., 2018, MNRAS, 475, 955
Rafieferantsoa M., Davé R., Anglés-Alcázar D., Katz N., Kollmeier J. A., Oppenheimer B. D., 2015, MNRAS, 453, 3980
Rahmati A., Pawlik A. H., Raičevic M., Schaye J., 2013, MNRAS, 430, 2427
Schmidt M., 1959, ApJ, 129, 243
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Springel V., 2005, MNRAS, 364, 1105
Stark D. V. et al., 2016, ApJ, 832, 126
Teimoorinia H., Ellison S. L., Patton D. R., 2017, MNRAS, 464, 3796
Vapnik V. N., 1995, The Nature of Statistical Learning Theory. Springer-Verlag, New York
Wang E., Wang J., Kauffmann G., Józsa G. I. G., Li C., 2015, MNRAS, 449, 2010
Wang J. et al., 2011, MNRAS, 412, 1081
Wang J. et al., 2013, MNRAS, 433, 270
York D. G. et al., 2000, AJ, 120, 1579
Zhang W., Li C., Kauffmann G., Zou H., Catinella B., Shen S., Guo Q., Chang R., 2009, MNRAS, 397, 1243

This paper has been typeset from a TEX/LATEX file prepared by the author.